Amendments to the Specification:

Please replace the paragraph beginning at page 11, line 14 with the following amended paragraph:

[0031] In one illustrative implementation, the new words analyzer 110 may implement different analyses depending on whether the Chinese input is a Chinese document or a user Chinese entry or query. With Chinese documents, the new words analyzer 110 may compute the frequency of each new character string and compute the coverage of subsets of new character strings to extract or determine a relatively small subset of new valid character strings that cover a sufficiently large portion of all new character strings found in the repository of Chinese documents. For a more complete analysis, the new words analyzer 110 may analyze all new character strings. Alternatively, the new words analyzer 110 may remove the new character strings with more than, for example, 7 (or other suitable number of) Chinese characters and group the remaining new character strings, i.e., those with 7 or fewer characters, according to the number of Chinese characters into 7 sets of new character strings (e.g., a separate group for strings having 7 contiguous characters, strings having 6 contiguous characters, strings having 5 contiguous characters, strings having 4 contiguous characters, strings having 3 contiguous characters, strings having 2 contiguous characters, and strings having 1 character). For each set of new character strings, the new words analyzer 110 may compute the coverage of its subsets. Specifically, the terms T in each of the 7 sets may be arranged in decreasing order of frequency $\{T_1, ..., T_n\}$. The coverage of a sub-list L_i, $\{T_1, ..., T_i\}$ is computed as the sum of the frequency of terms in the sub-list L_i divided by the sum of the frequency of all character strings in the set. Each of the 7 sets of new character strings may then be divided into three subsets where the first subset [[have]] has a coverage of greater than 98% and the first and second subsets have a combined coverage of greater than 99% within the set, for example. The character strings in the second subset may also be further evaluated manually to remove any unlikely character strings. The first subset and

the reduced second subset can be combined to form the new set of valid words generated from the repository of Chinese documents. These valid words are added to the dictionary.